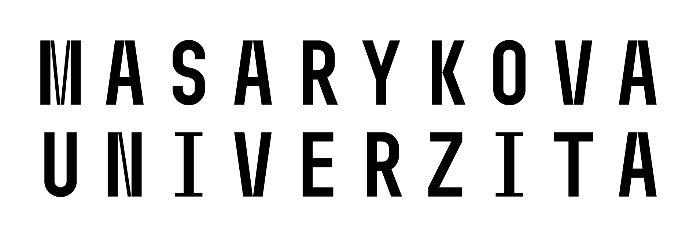


PŘÍRODOVĚDECKÁ FAKULTA

Diplomová práce

**Hoa Vu Thu**

**Brno 2020**



PŘÍRODOVĚDECKÁ FAKULTA

Klasifikace zdravotnických dat prostřednictvím neuronových sítí

Diplomová práce

**Hoa Vu Thu**

Vedoucí práce: RNDr. Martin Komenda, Ph.D.

Název ústavu

**Brno 2020**

**Bibliografický záznam**

|  |  |
| --- | --- |
| **Autor:** | Bc. Hoa Vu Thu  Přírodovědecká fakulta, Masarykova univerzita  Název ústavu |
| **Název práce:** | Klasifikace zdravotnických dat prostřednictvím neuronových sítí |
| **Studijní program:** | Experimentální biologie |
| **Studijní obor:** | Matematická biologie |
| **Vedoucí práce:** | RNDr. Martin Komenda, Ph.D. |
| **Akademický rok:** | 2019/2020 |
| **Počet stran:** | ??+?? |
| **Klíčová slova:** | Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo |

**Bibliographic Entry**

|  |  |
| --- | --- |
| **Author:** | Bc. Hoa Vu Thu  Faculty of Science, Masaryk University  Department of ... |
| **Title of Thesis:** | Health data classification using neural networks |
| **Degree programme:** | Experimental Biology |
| **Field of Study:** | Computational Biology |
| **Supervisor:** | RNDr. Martin Komenda, Ph.D. |
| **Academic Year:** | 2019/2020 |
| **Number of Pages:** | ??+?? |
| **Keywords:** | Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword |

**Abstrakt**

V této bakalářské/diplomové/rigorózní práci se věnujeme ...

100 – 2000 znakov bez medzier

**Abstract**

In this thesis we study ...

|  |
| --- |
| **Místo tohoto listu vložte kopii oficiálního zadání práce bez podpisů.** |

**Poděkování**

Na tomto místě bych chtěl(-a) poděkovat ...

**Prohlášení**

Prohlašuji, že jsem svoji bakalářskou/diplomovou/rigorózní práci vypracoval(‑a) samostatně s využitím informačních zdrojů, které jsou v práci citovány.

|  |  |
| --- | --- |
| Brno xx. měsíce 2020 | ………………………………  Hoa Vu Thu |

**Content**

[1. Introduction 9](#_Toc28971481)

[Czech national health system and web presentation 9](#_Toc28971482)

[Natural Language Processing 10](#_Toc28971483)

[Explanatory/Research questions 11](#_Toc28971484)

[2. Methods 12](#_Toc28971485)

[CRIPS-DM 12](#_Toc28971486)

[Natural language processing approaches and algorithms 13](#_Toc28971487)

[Neural networks 13](#_Toc28971488)

[Models for word representation 16](#_Toc28971489)

[Comparison of models 22](#_Toc28971490)

[Siamese architecture 23](#_Toc28971491)

[Technologies? 23](#_Toc28971492)

[Software 23](#_Toc28971493)

[Libraries? 23](#_Toc28971494)

[3. Results 25](#_Toc28971495)

[Data 25](#_Toc28971496)

[Data preprocessing 25](#_Toc28971497)

[Modeling 25](#_Toc28971498)

[Evaluation 25](#_Toc28971499)

[Comparison of models 26](#_Toc28971500)

[Deployment 26](#_Toc28971501)

[Evaluation 26](#_Toc28971502)

[4. Discussion 27](#_Toc28971503)

[5. Conclusion 28](#_Toc28971504)

[References 29](#_Toc28971505)

[Online references 30](#_Toc28971506)

# Introduction

## Czech national health system and web presentation

The Institute of Health Information and Statistics of the Czech Republic (IHIS) is an organisational component of the Czech Republic. It belongs to the Ministry of Health who delegated it to administrate the National Health Information System (NHIS). According to the Health Services Act the IHIS was specifically entrusted with the administration of:

* Data from the Program of statistical investigations of the Ministry of Health collected according to the Act on state statistical service
* National health registers including:
* the Czech National Cancer Registry
* the National register of hospitalisations
* the National register of reproduction health
* the National register of cardiovascular surgery and intervention
* the National register of joint replacement
* the National register of occupational diseases
* the National register of drug addict therapy
* the National register of injuries
* the National register of persons permanently excluded from blood donations
* the National register of autopsy and toxicological examination performed at forensic medicine departments
* the National register of health care providers
* the National register of health care professionals
* Data taken over from information systems on infectious diseases maintained according to Act on public health protection

All this data is used for monitoring health status of the population, the activity of health care providers, their economy and for obtaining the information about extend and quality of provided health services and for creation of health policy. NHIS is also designated for conducting and processing surveys on health status of population, on determinants, on the need and the consumption of health services, on its satisfaction and expenditure, for the needs of science and research in the field of the health.

The IHIS has currently several projects for example “Early detection of thyroid disease” in, or “Codification for Rare Disease”. For purpose of this thesis there will be introduced 2 of many projects and that is:

“**Methodological optimization and streamlining of the system of reimbursement of hospital care in the Czech Republic” -** this project was brought to the existence for the purpose of creating an information system of inpatient care reimbursement called CZ-DRG (diagnosis related group) and by its implementation there will be higher effectivity in management, controlling and optimization of distribution of financial flows in the health care.  
 The main goal is to build a long-term data, information and personnel base for the optimization and continuous cultivation of the system in the Czech Republic and to increase the predictive ability and effectiveness of reimbursement mechanisms for health care.  
 For achieving the main goal a Competence Centre and a representative network of reference hospitals had to be built and now it consists 49 hospitals, methodologies for evaluation of hospital cases, electronic and database tools for the implementation of the system had to be developed and verified in a representative network of reference hospitals, a new classification of hospitalization procedures had to be formed, the DRG classification system has to be redefined and a new coder’s manual had to be designed. The redefined DRG classification system can be found on <https://drg.uzis.cz/klasifikace-pripadu/web/>.

“**Centre for the Development of the Technology Platform of National Health Information System Registers, Modernization of their Content Extraction and Extension of their Information Capacity”** - the target of this project is methodological development of NHIS, increase its information capacity and value and also development of necessary supporting software applications and components. It will also contribute to more efficient collection of data, better usage of information by the state administration and overall development of the system of departmental registries NHIS.

## Natural Language Processing and text similarity

**Natural Language Processing (NLP) is an interdisciplinary area of research and application that investigates how computers understand and process natural language text or speech. One of the techniques of NLP is to measure text similarity that is important in text related research and applications in tasks, for example information retrieval, text classification, checking plagiarism etc. Finding similarity between words is an important part of text similarity which is then used as a primary stage for sentence, paragraph and document similarities.** There are 2 approaches how to measure text similarity:

* **Lexical similarity**
* **Semantic similarity**

**Lexical similarity** provides the similarity based on surface of text, so words are similar if they have a similar character sequence. To measure lexical similarity **string-based** algorithms are used which can be divided into **character-based** and **term-based** similarity measures.

**Semantic similarity** is similarity between texts is determined by their meaning, so words are similar semantically if they have the same thing, are opposite of each other, used in the same way, used in the same context and one is a type of another. Corpus-based and knowledge-based algorithms are used to measure semantic similarity. **Corpus-based similarity** computes similarity between words based on the information that is gained from large corpora. **Knowledge-based** similarity is used to determine the degree of similarity between words based on information derived from semantic networks. The most popular semantic network is WordNet which is a big lexical database of English language.

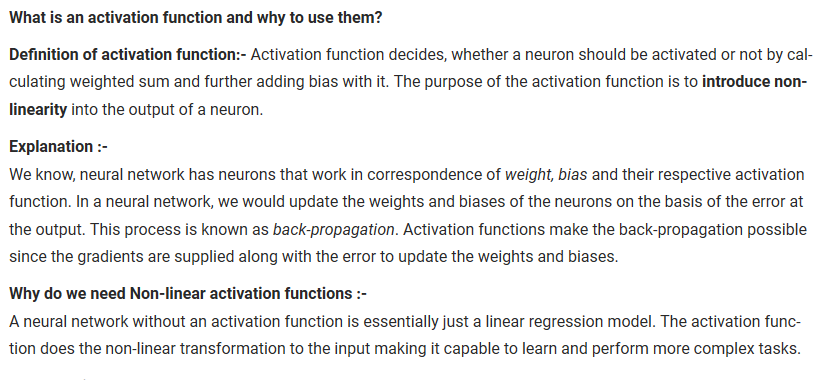
For the purpose of this diploma thesis there will be detailed description corpus-based similarity more specifically about word embeddings and deep learning methods on embeddings that use neural networks.

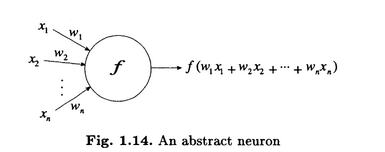
### Neural networks

Neural network (NN), also called artificial neural network (ANN) is a subset of machine learning algorithms. They are largely used in image-recognition. They can reach accuracy up to 90% and no other machine learning algorithms has better performance. In recent years they became quite popular also in natural language processing.

As it is mentioned above the NN is composed of units – neurons. Several units create a layer. There are 3 types of layers: input, output and hidden layer, which can have multiple layers. Between each layer are sets of weights. In general, the input units receive a signal from the outside of the NN, they process this signal and sent it to the follow-up units – the hidden units, which process the received signals in the same way as the input unit. This process is repeated within the hidden layer until the signals is received to the output units. It is assumed, that each unit provides additive contribution to the input of the unit which is connected. The input to the unit is linear combination of the weight and the output of the connected unit from previous layer which is connected to plus bias. …. About activation function decides whether a neuron will be activated or not. The purpose of the activation function is to introduce nonlinearity into the output of a neuron.

Activation function is a primitive function which transforms input of the unit into a defined output. ….





Blablabla

Three elements are important to define a neural network:

* Structure of the nodes
* The topology of the network
* The learning algorithm used to find the weights of the network

In the training process, every iteration consists of several steps:

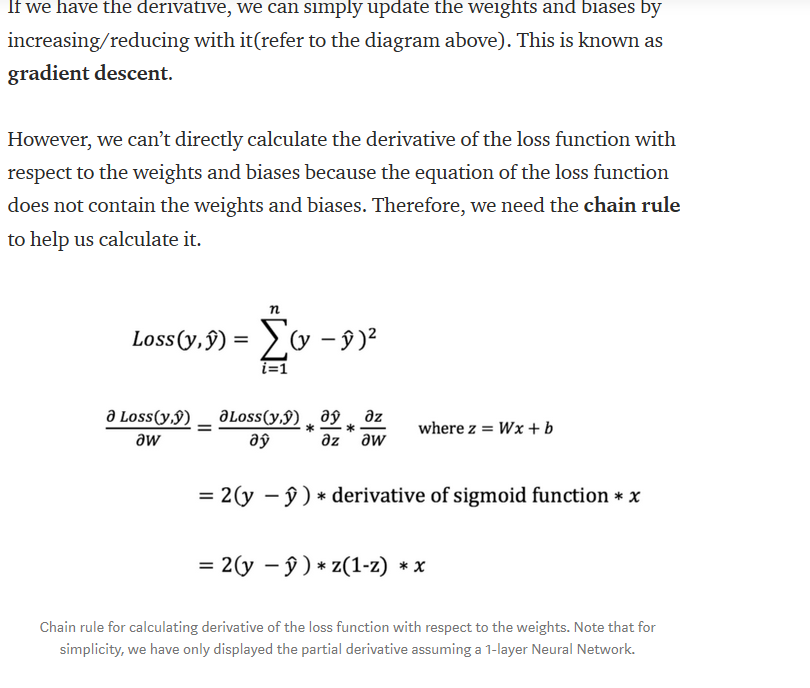
* Feedforward – prediction of the output
* Backpropagation – updating the weights and biases

**Loss function** is used to evaluate the goodness of the prediction from the trained neural network. There are many loss functions that can be used, for example the sum-of-squared error defined as:



Where y is … y is …. N is number of training examples. This is simply a sum of the differences between predicted value and the actual value.

The key in training neural network is to find the best set of weights and biases that minimizes the loss function but it is also important to avoid overtraining the network.

**Backpropagation -** blabalbal  
gradient descent – updating weights and biases by increasing or reducing with its derivative? 

[https://doi.org/10.1016/B978-0-12-741252-8.50010-8](https://doi.org/10.1016/B978-0-12-741252-8.50010-8" \o "Persistent link using digital object identifier" \t "_blank) - clanok, najst si h o doma

### Models for word representation

Distributed word representation = word vectors

<http://ruder.io/a-review-of-the-recent-history-of-nlp/>

distributional hypothesis – the idea that the meaning of a word is captured by the contexts on which it appears. Thus, the quality of the word vectors directly depends on the amount and quality of word vectors directly depends on the amount and quality of data they were trained on. (http://www.lrec-conf.org/proceedings/lrec2018/pdf/627.pdf)

**Reccurent neural network**

Sequential architecture

Sequence modeling like language modeling as it requires flexible modeling od context dependencies

**-Long short term memory (LSTM)**

**-Gated recurrent unit (GRU)**

**Models for Distributed representations of words**

Nowadays distributed word representations, which will be explained in this chapter, have been widely used in NLP. (http://www.lrec-conf.org/proceedings/lrec2018/pdf/627.pdf)

Sentence vector = paragraph vector = document vector – arbitrary-length abstract representation of the contextual meaning of a particular document type.

<https://arxiv.org/pdf/1405.4053.pdf>

Training, for successful application of distributed word representations it has to have a large corpora for training and using these pretrained models in downstream tasks

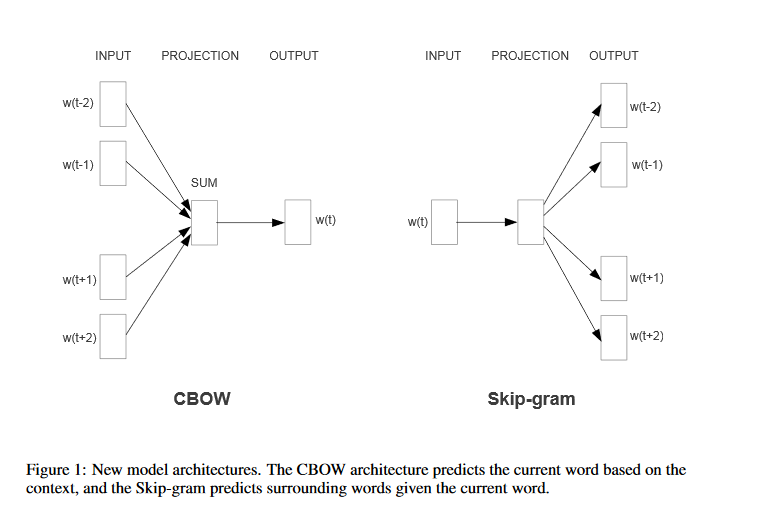
**Word2Vec**

Multiple sets of features for automatic genre classification of web documents no NN  
https://www.sciencedirect.com/science/article/abs/pii/S0306457304000676

Word2Vec are models used in natural language processing (NLP) to produce a word embedding. In general, word2vec is a two-layer neural network that is used to recreate linguistic context of words. As its name refer, word2vec transforms a corpus of text into a vector space with a big dimension and every word is assigned to a vector in the space. These vectors are called word vectors. Two words vector are close to each other when they share a similar context or meaning. Basically, it treats each words in corpus like an atomic entity and generate a vector for each word.

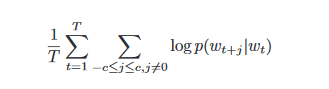
Word2Vec use two architecture: **continuous bag-of-word (CBOW) and continuous skip-gram**.

Continuous bag-of-words predicts the current word from the context. The order of context words does not affect prediction. In contrast, the continuous skip-gram uses the current word to predict the surrounding words and the weighs of words near to the current word are greater than the weights more distant from the current word. Continuous skip-gram is slower that CBOW but it has a better performance in case of infrequent words.

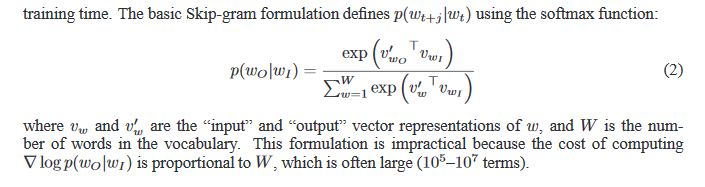


(<https://arxiv.org/pdf/1301.3781.pdf>)

**Skip-gram model** – this model search for word representations that help to predict the surrounding words in a sentence or document. Assume that we have a given sequence of training words *w1, w2, w3 … wT*. The skip-gram model maximizes the average log probability



Where *c* is the size of the training context (which can be a function of the center word *wt*). With the larger *c,* the higher accuracy can be obtained but the training time can be also increased.



(<https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>)

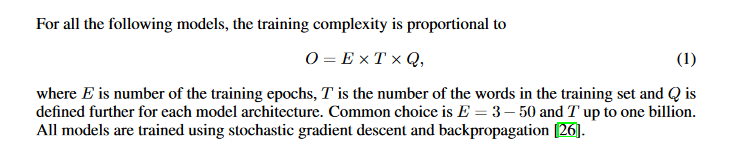
**Softmax**

**Hierarchical softmax**

**Negative sampling**

**Subsampling of frequent words**

The reason why is word2vec successful in word embeddings is poorly understood. The reasons for successful [word embedding](https://en.wikipedia.org/wiki/Word_embedding) learning in the word2vec framework are poorly understood. Goldberg and Levy point out that the word2vec objective function causes words that occur in similar contexts to have similar embeddings (as measured by [cosine similarity](https://en.wikipedia.org/wiki/Cosine_similarity)) and note that this is in line with J. R. Firth's [distributional hypothesis](https://en.wikipedia.org/wiki/Distributional_semantics). However, they note that this explanation is "very hand-wavy" and argue that a more formal explanation would be preferable.[[3]](https://en.wikipedia.org/wiki/Word2vec#cite_note-explain-3) Levy et al. (2015)[[16]](https://en.wikipedia.org/wiki/Word2vec#cite_note-16) show that much of the superior performance of word2vec or similar embeddings in downstream tasks is not a result of the models per se, but of the choice of specific hyperparameters. Transferring these hyperparameters to more 'traditional' approaches yields similar performances in downstream tasks. Arora et al (2016)[[17]](https://en.wikipedia.org/wiki/Word2vec#cite_note-17) explain word2vec and related algorithms as performing inference for a simple generative model for text, which involves a random walk generation process based upon loglinear topic model. They use this to explain some properties of word embeddings, including their use to solve analogies.  
The word embedding approach is able to capture multiple different degrees of similarity between words. Mikolov et al. (2013)[[18]](https://en.wikipedia.org/wiki/Word2vec#cite_note-18) found that semantic and syntactic patterns can be reproduced using vector arithmetic. Patterns such as “Man is to Woman as Brother is to Sister” can be generated through algebraic operations on the vector representations of these words such that the vector representation of “Brother” - ”Man” + ”Woman” produces a result which is closest to the vector representation of “Sister” in the model. Such relationships can be generated for a range of semantic relations (such as Country–Capital) as well as syntactic relations (e.g. present tense–past tense)



(https://arxiv.org/pdf/1301.3781.pdf)

**Doc2Vec**

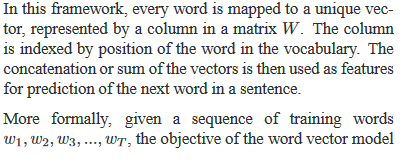
<https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=159497>

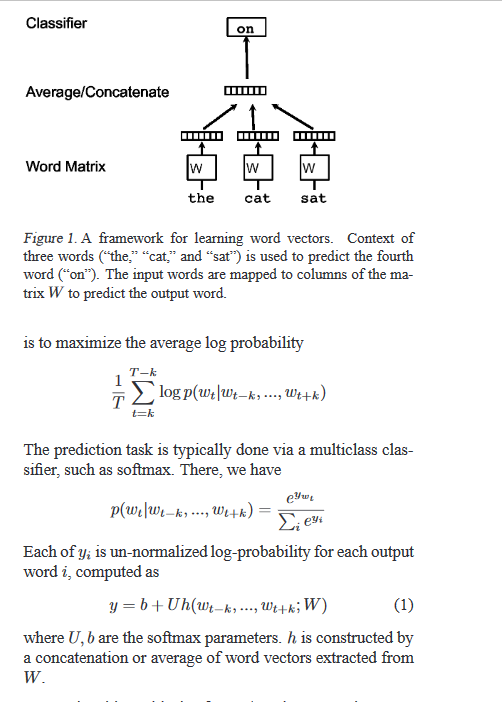
<https://towardsdatascience.com/multi-class-text-classification-with-doc2vec-logistic-regression-9da9947b43f4>

doi>[10.1145/1390156.1390177](https://doi.org/10.1145/1390156.1390177) A unified architecture for natural language processing: deep neural networks with multitask learning

Paragraph vector – unsupervised framework, it learns continuous distributed vector representation for pieces of texts (sentences or even documents). The vector representation is trained to predict words in a paragraph. The paragraph vector is concatenated with several word vectors from a paragraph and predict the following word in the given context. Words vectors and paragraph vectors, which are unique among paragraphs in contrast with word vectors, which are shared, are trained by stochastic gradient descent and backpropagation. At prediction time, the paragraph vectors are inferred by fixing the word vectors and training the new paragraph vector until convergence.

**Distributed vector representation of words**

the main goal is to predict a word given the other words in a context. 



http://ruder.io/word-embeddings-1/index.html

<https://medium.com/@dnberol/natural-language-processing-with-doc2vec-1d297059ac76>

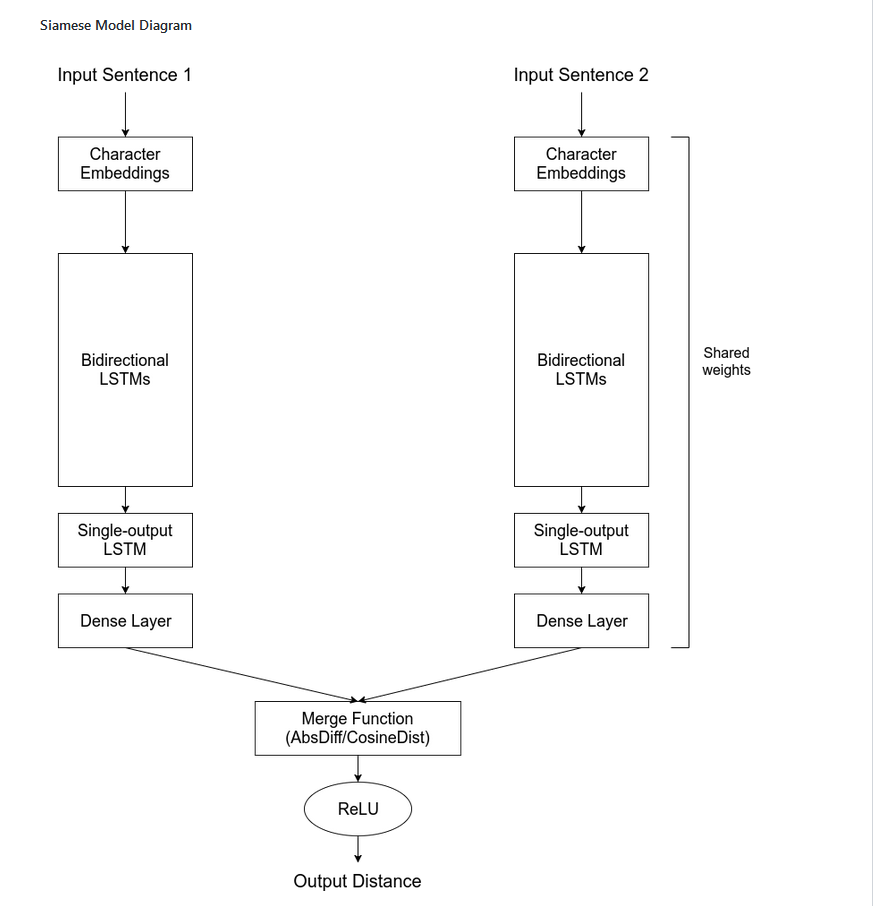
**GloVe**

Global vectors or Glove leverage global statistical information contained in a document. It is built on two main methods – global matrix factorization and local context windows. **Global matrix factorization** is used to reduce large term frequency matrices, which usually represent the occurrence or absence of words in a document. When it is applied to term frequency matrices it is called latent semantic analysis (LSA). Two methods are used in **local context windows:** CBOW and skip-gram that are mentioned in the chapter about word2vec.

**FastText**

Recently, deep neural networks have become very popular for text processing. These models have very good performance but they can be slow to train and test which limits using very large datasets. FastText uses a hierarchical classifier which reduces the time complexities of training and testing (from linear to logarithmic with respect to the number of classes). It also exploits the fact that class are imbalanced by using Huffman algorithm. FastText was developed by Facebook team.

Essentially it is an extension of word2vec but instead of considering whole words it considers sub-words. It treats each word as composed character of n-grams. The word vector is made of a sum of this character n grams. It generates better word embeddings for rare words. It can construct the vector for a word from its character n grams even if word does not appear in training corpus



<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

<http://colah.github.io/posts/2015-08-Understanding-LSTMs/?source=post_page-----37e2f46f1714---------------------->

<https://github.com/amansrivastava17/lstm-siamese-text-similarity>

doc2vec

<https://praveenbezawada.com/2018/01/25/document-similarity-using-gensim-dec2vec/>

**Neural networks for text comparison**

## Explanatory/Research questions

# Methods

## CRIPS-DM

This thesis was processed by a data mining method Cross Industry Standard process for Data Mining (CRISP-DM) which is illustrated in figure. This method is consisted of 6 following steps:

1. **Business understanding –** it is important to understand what is the main goal that is needed to be accomplished, define the outputs, design the project plan, check availability of resources, list requirements and consider pros and cons.
2. **Data understanding –** in this phase data are described and explored to understand them and their structure, also their quality must be verified.
3. **Data preparation –** this is the most time-consuming stage when it has to be decided on what data are going to be used and they need to be edited (cleaning, transformation, creating new attributes etc.) for further use.
4. **Modelling –** different models are applied, validated and optimized on data
5. **Evaluation –** results are evaluated, their quality is checked, and based on them the best model or models are chosen regarding to the prime goal
6. **Deployment –** create deployment, monitoring and maintenance plan and present the results to customer



Figure 1 Scheme of CRISP-DM

Order of these steps does not need to be kept. Output of one step is the input for following step. All these steps are connected mutually, and it is possible to go step back to make changes. The whole process has cyclic character for the reason that according to achieved results there can be more specific or even new research questions.

## Technologies

UDPipe, which is trainable pipeline for tokenization, lemmatization and dependency parsing of CoNLL-U files developed by Institute of Formal and Applied Linguistics at Charles University.

**Software**

Python 3., anaconda,

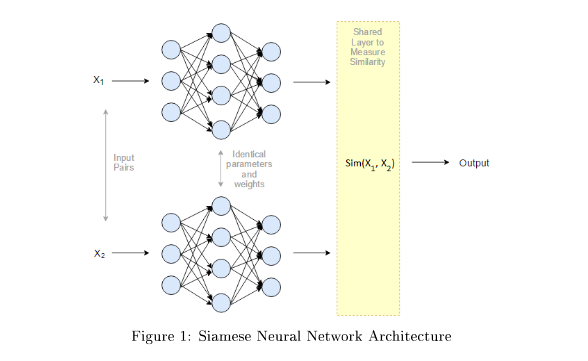
**Libraries?**

## Natural language processing approaches and algorithms

MALSTM

**Siamese architecture**

<http://ceur-ws.org/Vol-2028/paper8.pdf>



The objective of the siamese architecture is not to classify input images, but to differentiate between them.

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2789-2>

cosine of ht enagle between two feature vector is calculated and this represents the distance value

<https://arxiv.org/pdf/1904.11968.pdf>

# Results

In this chapter covers detailed description of used data, how these data were preprocessed, what methods were applied and their corresponding results according to CRIPS-DM method, except for the first step - business understanding. This step is already characterized in the “Explanatory/Research questions” chapter.

## Data

The corpus used for training models in this thesis was from (IDK) and for testing the model was from <https://drg.uzis.cz/klasifikace-pripadu/web/> and <https://reporting.uzis.cz/>. Those three websites are created by IHIS. There is information about the health care in the Czech Republic in Czech language.

The <https://drg.uzis.cz/klasifikace-pripadu/web/> is website representation of project “Methodological optimization and streamlining of the system of reimbursement of hospital care in the Czech republic”. This website is divided into five sections:

* **Introduction –** there are information about the project, data sources and contacts
* **Structure of system CZ-DRG –** in this section is description of each taxonomic level of system CZ-DRG and its labels
* **Definition lists –** there can be found classification rules for every hospitalization case and characterization of taxonomic level
* **Interactive classifier (grouper) –** in this section it is possible to classify hospitalization cases according to classification rules of system CZ-DRG
* **Analyses and publication –** this section provides published analytical outputs related to the CZ-DRG classification system or other detailed views of its taxonomic units.

Data were crawled from both website whereas from the first website were collected 1908 documents.

## Data preprocessing

All useful information from each XML documents were extracted from html tags using Xpath. Corpus was brought into existence by removing all numbers, special characters and stop words in data and by transforming all letters in words into lower case except for abbreviations such as HIV. Then the corpus was tokenized into words and lemmatized applying UDPipe.

## Modeling

For comparing similarity of two documents there were chosen two approaches:

* MaLSTM
* Doc2vec

## Evaluation

### Comparison of models

## Deployment

## Evaluation

# Discussion

# Conclusion

# References

# Online references

URL 1: <https://www.uzis.cz/> The Institute of Health Information and Statistics of the Czech republic [online] [cit. 23.November 2019]